DOCUMENT RESUME

ED 167 611 TM 008 424

AUTHOR	Love, Judith A.; And Others
TITLE	The Generalizability and Concurrent Validity of An Instrument to Describe College Teaching.
PUB DATE	Mar 78
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	*Classroom Observation Techniques; College Teachers; *Effective Teaching; Higher Education; Observation; *Student Evaluation of Teacher Performance; Teacher Behavior; *Teacher Evaluation; Teaching Skills; Test Construction; Test Reliability; *Test Validity
IDENTIFIERS	*Formal Lecture Observation Assessment Technique; *Generalizability Theory

ABSTRACT

The generalizability and concurrent validity of the FLOAT (Formal Lecture Observation Assessment Inventory) were studied. College instructors were rated, using the FLOAT, by their students and by trained observers who viewed videotapes of the classes. The teachers were also asked to evaluate themselves. A previous study of interrater reliability and student-observer agreement indicated that the instrument was reliable and that there were moderate to strong correlations between groups of raters. To study its generalizability, an analysis of variance was performed for the first fifteen FLOAT variables, used by three raters to rate nine instructors upon three occasions. Generalizability coefficients ranged from .66 to .96. Factor analysis of student ratings yielded two factors, one of which explained 61% of the rotated variance. Factor analysis of observer ratings produced similar results; the main factor, task orientation, explained most of the variance. Results of teacher self-ratings indicated a .42 correlation coefficient with student ratings, and .25 with observers. This difference was attributed to the observer's emphasis on category discrimination and to the students' perspective which did not distinguish greatly between teaching process and course content. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED167611

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE GENERALIZABILITY AND CONCURRENT VALIDITY OF AN INSTRUMENT TO DESCRIBE COLLEGE TEACHING

Judith A. Love
Teaching Resources Center

Jonathan Sandoval
Department of Education

Eli Cohen
Teaching Resources Center

University of California, Davis

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Judith Love

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM008 424

Paper presented at the annual meeting of
American Educational Research Association
Toronto, Ontario Session 25.23
March 27-31, 1978

Running Head: Generalizability and Concurrent Validity

© 1978
all rights reserved

Printed in U.S.A.

ABSTRACT

Innumerable systematic observation instruments are currently available to researchers wanting to observe and rate teaching behavior. Fewer are available to those who wish to study college teaching. The number further reduces to almost zero when measures with demonstrated reliability, generalizability and with concurrent validity are considered.

This paper describes an effort to establish the generalizability of a reliable observational instrument used to focus students' and observers' ratings of college lecturing. The rationale for such study is detailed, the methodology for testing the stability of the instrument's measures is described, and future plans for assessing its validity are outlined.

THE GENERALIZABILITY AND CONCURRENT VALIDITY OF AN INSTRUMENT TO DESCRIBE COLLEGE TEACHING

Descriptions of innumerable observational schedules pepper the literature and conference programs of the educational profession. Through their application, researchers purport to measure teacher effectiveness and, often, to isolate behaviors which contribute to greater or better student learning. Estimates of reliability of these instruments are bandied about and coefficients of inter-rater agreement are reported with much enthusiasm. There is little doubt that observers can be trained to observe and to rate such characteristics as "enthusiasm," "clarity" and "organization." There is also little information that these teacher characteristics make much difference in terms of student learning.

Because we need to study the teaching act in order to determine what does make a difference in terms of student learning, and because we need systematic observational instruments in order to do a defensible job of studying classroom activities, much research energy continues to be invested in the preparation and field-testing of such schedules.

In the first phase of this research (see Love-Dudek, Minnis, Sandoval and Yonge, 1977) we created a measure consisting of twenty-one variable commonly associated with "effective" lecturing. Table 1 presents the instrument (FLOAT).

Insert Table 1 about here

The inter-rater reliability estimates of four trained observers' ratings across sixty videotaped samples of university lecturing averaged .81. The ratings were factor analyzed and found to compose four discrete clusters or measures of lecturing behavior. Student ratings of the same lectures were also collected and were compared to the ratings of the trained observers. An examination of the correlations between these two groups of raters across the

sixty lectures indicated that the ratings of 18 of 21 variables were significantly correlated. The factor analysis of observers' ratings is presented in Table 2 and the correlations between students' and observers' ratings are presented in Table 3.

Insert Table 2 about here

Insert Table 3 about here

The encouraging reliability results led us to consider our phase one objective (to construct a reliable descriptive instrument for college teaching) achieved. The moderate to strong correlations among the two types of raters encouraged us to further explore the validity of the instrument's measures. Our current, phase two, goals became:

1. to establish the generalizability of the measures the FLOAT instrument provides;
2. to examine the factor breakdown of the students' ratings of the sixty lectures; and
3. to explore the nature of instructors' self-ratings in relationship to those of trained observers and those of students.

Inter-rater reliability does not imply stability of teacher behavior. The stability of teacher behavior must be estimated through the application of generalizability theory. Generalizability theory (Cronbach, Gleser, Nanda and Rajaratrum, 1972) permits the making of projections about the contribution of various research design components to the reliability of a behavioral measure. More specifically, the analysis of variance components, called facets, (in this case observers, occasions and instructors) are the conditions under which the sample observations are made. To the extent that ratings show limited variability across the facets of the design, an instrument's measure are said

to be generalizable. Limits to the generalizability of each measure come from constraints imposed by the conditions under which the observations and ratings were made, and from the estimation of error variance. Observers' ratings of instruction are the sum of "true" ratings plus an undifferentiated error term. Generalizability theory allows for the partitioning and attribution of those error terms to each of the facets of the data collection design. It is only then that a researcher can determine to what extent the sample s/he has studied identifies relationships in the unavailable universe.

Student evaluations of instruction differ, in nature, from trained observers' evaluations of instruction in that students approach the rating task with a dual orientation. The course content and the process are often confounded for students who seem to be unable to divorce one from the other. The trained observer, on the other hand, attends to content only as it contributes to process variables. Further, the trained observer approaches the task with the cultivated discrimination of a training (calibration) period. Thus, there are many forces at work lowering the concurrent validity of ratings of college lecturing behavior.

Self-ratings and self-reports have long been studied by social and counseling psychologists. There is good reason to believe that these measures are even more subject to error than are ratings by neutral observers. Motivation and self-esteem and their role in self-appraisal have been discussed by Combs (1965, 1972). The ability to apprehend the nature of the "successful" teaching act (Bess, 1977) contributes to the disparity between real and perceived behavior. These factors co-mingle to explain the error variance among those instructors who are more (and less) aware of their teaching competencies (and incompetencies).

Subjects

Videotapes of regularly-offered course-affiliated lectures were collected during the winter quarter of 1976, in the classrooms of 20 faculty and teaching assistants from nine departments at the University of California at Davis. Most requested videotaping as part of the U.C. Davis Teaching Resources Center's instructional improvement program. When informed of the proposed concurrent investigation, they agreed to contribute three videotapes and to complete the required paper work in return for detailed feedback from the Center's videotape analyst. Feedback sessions were held after the three videotaped samples of lecturing had been recorded in each instructor's classroom.

Observers and Observer Training

During a one-month period prior to the data-gathering stage, the authors of the FLOAT instrument met with two graduate students for a total of sixteen hours of training. The training period commenced with detailed discussions of the nature and scope of the categories for the observation, and culminated with three practice observation-rating sessions of three training tapes. When the observers each expressed confidence in their ability to observe and rate the teaching behaviors described by the instrument, the taping and rating segment of the study began.

Apparatus and Procedure

Details regarding these topics are available elsewhere. (Love-Dudek, et al., 1977.)

RESULTS AND DISCUSSION

Generalizability

An analysis of variance was performed for each of the first fifteen variables of the FLOAT instrument. The design fully crossed the effects of nine instructors, three raters, and three occasions, considering each as random. The analyses were based on this limited subset of the data and no multivariate analyses were performed due to computer limitations.

Generalizability coefficients were calculated in the manner explained by Erlich (1976). In brief, variance component estimates were calculated for instructor main effect, and all the interaction effects which include the instructor effect, by using the Analysis of Variance Table and expected mean square formuli, for each variable.

The generalizability coefficient, σ^2, is the ratio of the universe score to the observed score variances. For our design, the universe score variance is instructor variance and the observed score variance is composed of the variance of the instructors, plus the variance components for each of the interaction effects which involve the instructor effect.

Symbolically,

$$\hat{\rho}^2 = \frac{\sigma^2_t}{\sigma^2 t + \frac{\sigma^2 ti}{k} + \frac{\sigma^2 tj}{n} + \frac{\sigma^2 tij,e}{kn}}$$

where t is the instructor index

i is the occasion index

j is the observer index

k is the number of occasions (3)

and n is the number of observers (3)

The results of the analysis are shown in Table 4.

Insert Table 4 about here

The magnitudes of the generalizability coefficients are high, providing evidence for the assertion that FLOAT is a reliable and generalizable instrument, across trained raters and across occasions.

A cautionary note is in order. In the classroom situation, course material differs from day to day, and from course to course. The instructor tends to be inconsistent in his/her responses in different contexts. Similar qualifications have been made by Erlich (1976), Gillmore, Kane, and Naccarto (1978), Erick and Semmel (1974), and Kane and Brennan (1977).

Further, it may be argued that both the sample of instructors and the raters were nonrandom. The instructors were volunteers and two of the four raters were the creators of the instrument. Cornfield (1956) discusses some effects of sample constraints on results.

Lastly, as pointed out by Shavelson and Atwood, situational and observational facets must be systematically varied to test the effect of a particular facet on stability (1977). The situational facets of raters and occasions have been accounted for while the observational facets of period and of type of rating scale have not. We concur with Shavelson and Dempsey's recommendation that a systematic effort should be made to resolve the issues of conceptualization and measurement of teacher effectiveness (1976).

Factor Analysis of Students' Ratings

In phase one (Love-Dudek et al., 1977), we found high agreement between students' and observers' ratings of the 60 videotaped lectures (Table 3). These results lead us to explore the structure of the students' ratings, through factor analysis, to see if the factors themselves were similar. The factor analyses of the observers' ratings, reported in 1977 and discussed earlier in this paper, produced one primary factor and three others of lesser importance (Table 2). A factor analysis of the students' ratings yielded two factors after rotation, one explaining a majority of the variance and a minor factor, explaining only a small amount of the variance (Table 5).

---

insert Table 5 about here

The two factor analyses (of observers' ratings, in Table 2 and of students' ratings, in Table 5) do indeed have much in common. Most of the items with high loading on the observer primary factor also have high loading on the student primary factor. The items which did not load highly on the observer primary factor (items 8 through 12) loaded on a second factor for students' ratings accounting for just 5% of the variance and on the second, third, and fourth factors of the observers' factor analyses (accounting collectively for just 25% of the variance). The one main observer factor, which we called Task-Orientation, explains most of the variance. The other factor(s) represent interactive sequences, such as support behaviors and questioning skills. Again, these factors are of minor importance.

Thirty-five percent of the variance in student ratings must be relegated to unexplained error. The observer ratings must be considered more discriminating.

That the raters judgements yielded several factors was probably due to their efforts to categorize behaviors and not rate by global impressions. As mentioned earlier, the relationship of content to process confounds the ratings. Another real limit to accurate measurement is the students' exposure to an entire course contrasted to the observers' exposure to just three lectures. Whitely and Doyle (1976) raised a new question: do students approach the rating task with inherently different clusters of behaviors ("implicit factors") in

mind? That is, rather than permitting the natural variability of instructors' behaviors to contribute to factor formation, do students (or observers) apply rating instruments to teaching acts in a manner that accentuates an internalized "covariance"? We intended to address this issue in this paper, but felt a thorough treatment was precluded by the absence of the pre-evaluation sorting task suggested by Whitely and Doyle (1976).

Self-Evaluations

As previously mentioned, instructors were asked to complete an evaluation of their own lecturing behaviors. The evaluation was based on the items of the FLOAT instrument (Table 1), but written in the first-person, singular. A nine-point rating scale was used. Pearson product-moment correlations were computed for the mean ratings of students x observers, observers x instructors and students x instructors. The students and observers evidenced moderately strong agreement ($r = .50$), considering the content-process orientation. However, a major discrepancy existed between the students x instructors and the observers x instructors: the former coefficient was .42, while the latter was .25. The highly significant difference between the two relationships must be attributed to, again, 1) the emphasis on category discrimination during observer training and 2) the content-process problem and its relationship to perspective.

Plans for Future Research

Early in 1976, as our three-phase research plan was outlined, we saw as our third goal an examination of external validity (see Crawford, Gage and Stallings, 1977) for a discussion of internal vs. external validity). Our plans to meet this goal would involve us in research on teacher effects (the causal relationship between the behaviors measured by FLOAT and student outcomes).

Gage (1972) anticipated many of the problems that would emerge in these process-product studies of teaching; particularly the dilemma created by the interactive nature of the teaching and the learning processes. Shavelson and Dempsey (1976) provided the criterion-of-effectiveness paradigm, which is also student-outcome oriented, but condemned both paradigms as yielding unstable results. It was these set of issues which we planned to incorporate into the third stage of our research plans.

What should we do next? Several additional issues have captured our interest. We intend to revise our instrument in view of the implicit factors of students (Whitely & Doyle, 1976). Also, we intend to include a situation-specific context in our next generalizability design (Calkins, Borich, Pascone and Kugle, 1977). Further, we intend to explore a Bayesian interpretation of the rating process. Specifically, there is reason to believe that raters may get an impression of an entire lecture on the basis of the first few minutes. If that is so, then the observational facet of the five-minute periods should be included in our generalizability formula as a separate facet. We intend to explore that possibility in future research.

Table 1 - FLOAT*

A list of teaching behaviors employed in the college classroom and intended to structure the systematic observation and rating of college teaching.

Item	Variable	Description
1	$Clarity_1$	The instructor speaks audibly and clearly.
2	$Clarity_2$	Instructor is able to simplify concepts.
3	$Enthusiasm_1$	Instructor is energetic, stimulating, involved.
4	$Enthusiasm_2$	Instructor appears to enjoy both content and process.
5	$Organization\text{-}Structuring_1$	Instructor presents materials in an orderly, logical manner, making clear and logical transitions.
6	$Organization\text{-}Structuring_2$	Instructor integrates, relates and categorizes to aid both comprehension and note-taking.
7	$Knowledge\ of\ Subject\ Matter_1$	Instructor projects command of material.
8	$Knowledge\ of\ Subject\ Matter_2$	Instructor answers student questions and concerns or redirects them to appropriate resources.
9	$Interaction_1$	Instructor provides opportunity for expression of other opinions, assenting and dissenting.
10	$Interaction_2$	Instructor uses student ideas and comments.
11	$Support_1$	Instructor solicits and receives involvement in an accepting, non-threatening manner.
12	$Support_2$	Instructor gives feedback appropriately.
13	$Questioning\ Tactics_1$	Instructor displays facility with strategic questioning tactics.
14	$Questioning\ Tactics_2$	Instructor utilizes questions for many purposes.
15	$Questioning\ Tactics_3$	Instructor addresses questions to many individuals and groups.
16	$Variability_1$	Instructor displays a repertoire of skills, adapting his pace and techniques to the material being presented.
17	$Variability_2$	Instructor utilizes diverse resources to illustrate principles
18	$Stage\ Presence_1$	Instructor's verbal behavior is confident and effective.
19	$Stage\ Presence_2$	Instructor's physical demeanor is attractive and assured.
20	$Use\ of\ Visuals_1$	Media easy to follow and pertinent to lecture.
21	$Use\ of\ Visuals_2$	Charts, diagrams, boardwork are appropriate and presentation is effective.

*Formal Lecture Observation Assessment Technique

© 1976

TABLE 2

Ratings of College Teaching

Varimax Rotated Factor Matrix by Grouped Observers

Factor 1
Task Orientation

Item	Description	Coefficient
2	"...simplification of concepts..."	.87
3	"...energetic, stimulating, involved..."	.85
4	"...enjoys content and process..."	.84
5	"...orderly, logical presentation..."	.93
6	"...integrates, relates, categorizes..."	.92
7	"...projects command..."	.94
8	"...answers or redirects questions..."	.69
10	"...uses student ideas..."	.65
11	"...solicits involvement..."	.51
12	"...gives feedback..."	.56
16	"...repertoire of skills..."	.73
17	"...uses diverse resources..."	.74
18	"...confident, effective verbal behavior..."	.64
	Eigenvalue	12.398
	Percent of Rotated Variance	74.4

Factor 2
Use of Resources

Item	Description	Coefficient
4	"...enjoys content and process..."	.46
9	"...opportunity for opinion..."	.48
10	"...uses ideas..."	.47
11	"...solicits involvement..."	.71
12	"...gives feedback..."	.64
13	"...strategic question..."	.91
14	"...many purposes for questions..."	.71
15	"...questions many individuals..."	.90
	Eigenvalue	1.922
	Per cent of Rotated Variance	11.5

Factor 3
Verbal Presence

Item	Description	Coefficient
1	"...speaks audibly..."	.77
8	"...answers or redirects questions..."	.38
9	"...opportunity for opinions..."	.40
16	"...repertoire of skills..."	.43
17	"...utilizes diverse resources..."	.51
18	"...confident, effective verbal behavior..."	.58
	Eigenvalue	1.356
	Per cent of Rotated Variance	8.1

Factor 4
Physical Presence

Item	Description	Coefficient
19	"...attractive physical demeanor.."	.80
21	"...use of blackboard..."	.87
	Eigenvalue	0.978
	Per cent of Rotated Variance	5.9

TABLE 3

Correlations between Grouped Students and Grouped Trained Observers' Mean Ratings of Lecturing Behavior across Three Samples of Twenty Instructors' Lectures for each of Twenty-one Variables.

Item	Variables	Correlations	Item	Variables	Correlations
1	Clarity$_1$	.88**	11	Support$_1$	.57**
2	Clarity$_2$	.59**	12	Support$_2$	.51**
3	Enthusiasm$_1$	.67**	13	Questioning$_1$	.48*
4	Enthusiasm$_2$	.59**	14	Questioning$_2$	.41*
5	Organization-Structuring$_1$	.43*	15	Questioning$_3$	.52**
6	Organization Structuring$_2$	.33	16	Variability$_1$	.43*
7	Knowledge of Subject$_1$	.44*	17	Variability$_2$	.52.**
8	Knowledge of Subject$_2$	.49**	18	Stage Presence$_1$	.49**
9	Interaction$_1$	.43*	19	Stage Presence$_2$	.02
10	Interaction$_2$	.47*	20	Use of Visuals$_1$	.49**
			21	Use of Visuals$_2$	.37

**p ≤ .01
*p ≤ .05

TABLE 4

Generalizability Coefficients Based on an Analysis of Variance Among Three Observers' Ratings of College Lecturing

Variable		Generalizability Coefficient
01	$Clarity_1$	.84
02	$Clarity_2$	.96
03	$Enthusiasm_1$	.87
04	$Enthusiasm_2$	.96
05	$Organization\ Structuring_1$	.96
06	$Organization\ Structuring_2$	.94
07	$Knowledge\ of\ Subject\ Matter_1$	.91
08	$Knowledge\ of\ Subject\ Matter_2$	.79
09	$Interaction_1$	.89
10	$Interaction_2$	.75
11	$Support_1$	.88
12	$Support_2$	.92
13	$Questioning\ Tactics_1$	.87
14	$Questioning\ Tactics_2$	.66
15	$Questioning\ Tactics_3$	.88

TABLE 5

Factor Analysis of 1355 Students' Ratings of 60 College Lectures Across Twenty-One Characteristics

Item	Factor 1	Factor 2
01 $Clarity_1$	.72	.01
02 $Clarity_2$	.87	- .05
03 $Enthusiasm_1$	.68	.14
04 $Enthusiasm_2$	.58	.23
05 $Organization\ Structuring_1$	.95	- .14
06 $Organization\ Structuring_2$	.88	- .06
07 $Knowledge\ of\ Subject\ Matter_1$	.78	.04
08 $Knowledge\ of\ Subject\ Matter_2$	.39	.43
09 $Interaction_1$	.09	.88
10 $Interaction_2$	.01	.84
11 $Support_1$	.02	.80
12 $Support_2$	.28	.63
13 $Questioning\ Tactics_1$	.31	.56
14 $Questioning\ Tactics_2$	.23	.64
15 $Questioning\ Tactics_3$	.00	.76
16 $Variability_1$	.61	.27
17 $Variability_2$	.44	.38
18 $Stage\ Presence_1$	.77	.07
19 $Stage\ Presence_2$	.74	.05
20 $Use\ of\ Visuals_1$	-no data-	-no data-
21 $Use\ of\ Visuals_2$	.72	.06
	61% of rotated variance	5% of rotated variance

BIBLIOGRAPHY

Bess, J.L. The motivation to teach. Journal of Higher Education, Vol. 48, No. 3 May/June, 1977, 243-258.

Brumet, M.E. Bayesian inferential processes in educational evaluation. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, March, 1976.

Calkins, D.S., Borich, G.D., Pascone, M. and Kugle, C.L. Generalizability of teacher behaviors across classroom observation systems. Austin: Research and Development Center for Teacher Education, 1977.

Combs, A.W. The professional education of teachers: A perceptual view of teacher education. Boston: Allyn & Bacon, 1965.

Combs, A. W. Some basic concepts for teacher educators. Journal of Teacher Education, 1972, 23 (3), 286-290.

Cornfield, J. and Tukey, J.W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Crawford, J., Gage, N.L., and Stallings, J.A. Methods for maximizing the validity of experiments on teaching. Paper presented at the annual meeting of the American Educational Research Association, April, 1977.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnum, N., 1972. The dependability of behavioral measures: the theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc., 1972.

Erlich, D. A study of the generalizability of measures of teacher behavior. Unpublished dissertation, University of California, Los Angeles, 1976.

Frick, T. and Semmel, M.I. Observational records: observer agreement and reliabilities. Bloomington, Indiana: Center for Innovation in Teaching the Handicapped, 1974.

Gage, N.L. Teacher effectiveness and teacher education: the search for a scientific basis. Palo Alto: Pacific Books, 1972.

Gillmore, G.M., Kane, M.T., Naccarato, R.W. The generalizability of student ratings of instruction: estimation of the teacher and course components. Journal of Educational Measurement Vol. 15, No. 1, Spring, 1978.

Kane, M.T. and Brennan, R.L. The generalizability of class means. Review of Educational Research, Spring, 1977, Vol. 47, No. 1, 267-292.

Love-Dudek, J.A., Minnis,D.L., Sandoval, J.H. and Yonge, G.D. An instrument to describe college teaching: a reliability study. Paper presented at the annual meeting of the American Educational Research Association. New York, April, 1977.

Shavelson, R. and Dempsey, N. Generalizability of measures of teaching behavior. Review of Educational Research, Vol. 46 (No. 4), Fall, 1976.

Shavelson, R. and Atwood, N. Generalizability of measures of teaching process. In Borich, G.D. (Ed.) The appraisal of teaching: concepts and process. Reading, Ma: Addison-Wesley, 1977.

Whitely, S.E. and Doyle, K.O., Implicit theories in student ratings. American Educational Research Journal Vol. 13 (4) Fall, 1976, 241-253.